Making Speech Synthesis More Accessible to Older People

Maria Wolters[1], Pauline Campbell[2], Christine DePlacido[2], Amy Liddell[2], David Owens[2]

[1]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
[2]Audiology Division, Queen Margaret University, Edinburgh, UK
mwolters@inf.ed.ac.uk, (pcampbell|cdeplacido|06006484|06005471@qmu.ac.uk)

Abstract

In this paper, we report on an experiment that tested users' ability to understand the content of spoken auditory reminders. Users heard meeting reminders and medication reminders spoken in both a natural and a synthetic voice. Our results show that older users can understand synthetic speech as well as younger users provided that the prompt texts are well-designed, using familiar words and contextual cues. As soon as unfamiliar and complex words are introduced, users' hearing affects how well they can understand the synthetic voice, even if their hearing would pass common screening tests for speech synthesis experiments. Although hearing thresholds correlate best with users' performance, central auditory processing may also influence performance, especially when complex errors are made.

1. Introduction

Older people are a key user group for speech synthesisers. Not only is the percentage of older people in the population increasing, but there are also many groups of older people who will clearly benefit from voice interfaces. Take for example people whose arthritis restricts the motion of their arms and hands: This user group will find it very difficult to navigate traditional graphical user interfaces. Moreover, as the baby boomer generation enters old age, older people are becoming more familiar with and amenable to using computer technology. But there is a fly in the ointment: Older people are also far more likely to have hearing problems than younger users. However, we should be able to optimise our synthetic voices to help compensate for these problems. To achieve this, we need to understand what makes synthetic speech more difficult to understand for older people. In this paper, we report a detailed error analysis of an intelligibility experiment that potentially hints at the direction to take. After a short review of the literature (Section 2), we describe the assessment battery each participant underwent (Sections 3.2 and 3.3) and the experiment itself (Section 3.4). In Section 4, we relate error patterns to selected aspects of participants' hearing, participants' cognitive ability, and problems with the synthetic stimuli. Finally, in Section 5, we suggest how synthesis systems might address the issues found.

2. Background

Older listeners have problems understanding synthetic speech, in particular if they have hearing problems [1], and if there are no contextual cues to compensate for the diminished acoustic cues [2]. Unfortunately, most of the research investigating potential reasons for these problems has not been carried out on unit-selection voices, but on formant synthesisers. The two major problems with formant synthesisers are the dearth of acoustic information in the signal [3] and incorrect prosody [4]. These problems with decoding the signal may place a higher cognitive load on listeners [5]. This increased load may affect older listeners more than younger ones [6]. Since concatenative approaches preserve far more of the acoustic signal than formant synthesisers, dearth of information should not be a problem anymore. Instead, we have problems with spectral mismatches at joins between units, spectral distortion due to signal processing, and temporal distortion due to wrong durations. It is central auditory processing mechanisms that are responsible for tasks such as detecting gaps or compensating for spectral and temporal distortions. Problems with central auditory processing are not picked up by standard pure-tone audiometry. Therefore, we need to expand our range of measures.

The results of Roring et al. [2] may suggest that we need to be particularly careful not to introduce distortions due to signal processing. Their stimuli were generated using an American English diphone voice as supplied with the open source version of Festival [7]. Stimuli were presented at two rates, normal (210 words-per-minute (wpm), duration parameter 1.0), and slow (150 wpm, duration parameter 1.5). The slow rate was chosen based on a 1995 study of DECtalk [8]. Older adults performed significantly worse at the slower rate, which was generated by setting Festival's duration parameter to 1.5 instead of 1.0. Not having heard the original stimuli, we can only speculate that this result was due to increased distortions introduced by PSOLA. As older adults are less able to compensate for those distortions than younger adults, this may partly explain the finding. Langner and Black [1] compared, among other options, speech that was recorded while the speaker was listening to time-varying noise (speech-in-noise) and synthetic speech that was post-filtered to mimic the spectral characteristics speech-in-noise. The original speech-in-noise had a positive effect on performance, the filtered version did not.

Although both Roring et al. [2] and Langner and Black [1] examine the role of hearing problems, neither was able to perform a comprehensive hearing assessment of their participants. Langner and Black relied on self-reports of hearing problems, while Roring et al. used pure-tone audiometry to determine participants' hearing threshold, averaging thresholds for 0.25, 0.5, 1, 2, 3, 4, 6, and 8 kHz.

Roring et al. concluded from their study of the Festival diphone voice, published in 2007, that "[s]ynthetic speech fidelity must be improved significantly before becoming truly useful for the older adult population." [2, p. 25]. One of the aims of this paper and its companion paper [9] is to assess whether unit selection has delivered this significant improvement. In our previous analyses of the data set reported on here [9], we examined correlations between pure-tone thresholds and intelligibility in more detail. We found that the most important threshold to consider is the average threshold for 1, 2, and 3 kHz, corresponding to the range of F2. We also noticed that extended

high frequency (UHF) thresholds above 9 kHz correlated well with participants' performance. UHF thresholds are a potential indicator of the general health of the cochlea, since hearing loss begins at the highest frequencies of 20 kHz and propagates down with age. These correlations were not due to a subset of participants with particularly pathological hearing—we can see these trends even in participants who would pass standard screening tests where 0.5, 1, 2, and 4 kHz pure tones are presented at 20dB.

From this brief review of the literature, we see that we know very little about the way in which age-related changes in hearing affect the intelligibility of synthetic speech, in particular unit selection. These age-related changes do not necessarily have to be pathological to affect a person's performance. Furthermore, the role of central auditory processing has barely been explored, even though it is key to compensating for artefacts introduced during the synthesis process.

3. Experiment

3.1. Participants

44 participants took part in our experiment. 12 were aged between 20 and 30, 20 between 50 and 60, and 12 between 60 and 70. The 20-30 group served as controls who showed very few signs of auditory ageing. The 50-60 group were included because they are more likely to show clear evidence of auditory ageing, but less likely to have complex pathologies or require a hearing aid. Finally, the 60-70 group fits with the type of participants that are typically labelled "older". We pooled the participants aged between 50-70 into a generic "older" group because chronological age is notoriously bad at predicting changes in ability [10].

3.2. Cognitive Assessments

We used the Prospective and Retrospective Memory Questionnaire [11] to screen for major memory problems. All scores were well within the normal range. In addition, all participants completed a working memory span (WMS) test [12] that was scored from an answer sheet. The test was presented visually because auditory presentation might affect scores [13]. We used WMS because the experimental task involved remembering the information presented in reminders (cf. Section 3.4 for more detail), and because WMS is highly correlated with other measures of cognitive functioning [10]. Older participants had a significantly lower WMS than younger participants (t-test,t=5.33,df=29.606,p$<$0.00001). The 20-30's scored on average 38 points out of 42, the 50-70's scored 27. The spread of scores in our test is considerable, with 25% of all participants scoring 24 of 42 possible points or less.

3.3. Audiological Assessments

3.3.1. Pure-Tone Audiometry

Pure-tone (PTA) and ultra high-frequency (UHF) audiometry was measured on a recently calibrated audiometer (Grason-Stadler, Milford, NH; model GSI 61) in a double-walled soundproofed room (Industrial Acoustics Corporation, Staines, Middlesex, UK). Air-conduction thresholds were measured for each ear at 0.25, 0.5, 1, 2, 3, 4, 6, and 8 kHz following the procedure recommended by the British Society of Audiology [14]. UHF thresholds were established at 9, 10, 11.2, 12.5, 14, 16, 18, and 20 kHz. If a participant was unable to detect a tone at the loudest setting `IntMax` for that particular frequency, their threshold for that frequency was recorded as `IntMax` + 5 dB. Testing always began with the better ear in all subjects. Since there are significant differences between the two ears, data from the right and the left ear will be reported separately in this analysis. In this paper, we use the following thresholds:

Trad: Average of 0.5, 1, 2, and 4 kHz, the frequencies conventionally used for screening participants in speech synthesis experiments

F2: Average of 1, 2, and 3 kHz, the frequency range of F2, which has been found to correlate with participants' ability to understand synthetic speech [9]

UHF: Average of 9, 10, 11.2, 12.5, 14, 16, 18, and 20 kHz

3.3.2. Gap Detection

The aim of the gap detection test is to establish the smallest gap between two carrier stimuli that participants can detect. Instead of psychoacoustic testing procedures, we used the Random Gap Detection Test [15], which samples gap detection ability at a fixed set of seven intervals, namely 0, 2, 5, 10, 15, 20, 25, 30, and 40 ms. The sequence in which these intervals are presented is randomised. The stimuli consisted of a 1000 Hz calibration tone and two subtests, the first covering the four frequencies 0.5, 1, 2, and 4 kHz, the second covering clicks. In this paper, we only report results for clicks, because we did not find any correlations between participants' performance on the synthetic speech test and their ability to detect a gap between two tones [16]. This finding is mirrored by studies which found that people's ability to detect gaps between tones does not correlate well with their ability to understand speech in noise, while their ability to detect gaps in noise does [17]. All test items were presented binaurally through a GSI 61 audiometer (model GSI 61; Grason-Stadler, Milford, NH) and a high fidelity Sony cassette with calibrated TDH-49 headphones.

3.3.3. Speech Audiometry

The speech audiometry test used a set of 20 standard CVC word lists [18]. Each list was 10 words long. After each word, participants need to repeat what they heard. The score is the number of phonemes that were repeated correctly, with the maximum score 30 (10 words $\times$ 3 phonemes). Word lists were initially presented at a comfortable intensity derived from participants' PTA scores. That intensity was increased until participants scored 30 out of 30 phonemes correct, and then lowered again until participants' score dropped to 3 out of 30 phonemes (10%) or worse. Intensity was changed by 5 dB at a time.

3.4. Synthesis Experiment

For this study, we used stimuli that are closely modelled on a real-life application—task reminders. Task reminders were chosen because they are an integral part of many relevant applications, ranging from electronic diaries to cognitive prosthetics [19]. Since our research focusses on adapting speech technology to the home care domain, we investigated two relevant types of reminders: reminders to meet a specific person at a given time, and reminders to take a specific medication at a given time. 32 reminders were generated, 16 meeting reminders and 16 medication reminders. Time preceded person or medication in half the sentences, person/medication preceded time in the other half. Table 1 shows the sentence templates that were used. Each template was used eight times.

Table 1: *Reminder Templates*

Reminder	Template
Meeting	At TIME, you are meeting PERSON.
	You are meeting PERSON at TIME.
Medication	At TIME, you need to take your MEDICATION.
	You need to take your MEDICATION at TIME.

3.4.1. Stimuli

There were three categories of target stimuli, times (easiest), person names (medium difficulty), and medication names (most difficult). Since *temporal expressions* are relatively distinct from each other, it is difficult to elicit errors. We addressed this problem by focussing on two sets of phonologically similar hours: "seven", "eleven" and "twelve" and "one", "nine", and "ten". We added further complexity by adding complex expressions such as X to HOUR and X past HOUR, where X was one of "ten", "twenty", and "a quarter". We chose *proper names* that matched the pattern C_1VC_2, where both consonants were oral or nasal stops, because stops are more easily confundable than other consonant types [20, 21]. For each proper name (except for "Dan"), we ensured that there was at least one other proper name that differed from the name by just one consonant. *Medication names* were constructed by recombining morphemes taken from actual medication names. Care was taken to ensure that the medication names did not resemble any existing or commonly used medication to avoid familiarity effects. All names are 3-4 syllables long; seven contain at least one consonant cluster. Table 2 lists all targets used in the experiment.

3.4.2. Voices

For the *synthetic speech* condition, all 32 reminders were synthesised using Scottish female voice "Heather" of the unit selection speech synthesis system Cerevoice [22]. Medication names were added to the lexicon before synthesis to eliminate problems due to letter-to-sound rules. The transcriptions were adjusted to render them maximally intelligible. No other aspects of the synthetic speech were adjusted.

For the *natural speech* condition, the reminders were read by the same speaker who provided the source material for the synthetic voice. The natural speech was then postprocessed using the procedures used for creating synthetic speech: high-pass filtering with a cut-off frequency of 70 Hz, then downsampling to 16kHz, and finally encoding and decoding with the tools `speexenc` and `speexdec`. This procedure ensures an exceptionally close matching between human and synthetic speech.

3.4.3. Experiment Design and Procedure

Four stimulus lists were created, each comprising 32 reminders. Each reminder was followed by a short question, recorded using the same natural voice as that used for the reminders. Each participant only heard one of the four lists. Each reminder was presented using the synthetic voice in two lists, and using natural speech in the remaining two. In two lists (one synthetic, one natural), participants were asked for the first item of a given reminder, while in the other two conditions, participants were asked for the second item.

The sequence of reminders was randomised once and then kept constant for all four lists. Each participant had to correctly remember 32 targets: 8 times presented in a natural voice, 8 times presented in a synthetic voice, 4 medication names presented in a synthetic voice, 4 medication names presented using a human voice, 4 person names presented using a human voice, and 4 person names using a synthetic voice.

Participants replied verbally with the information which they had been asked to recall. All responses were written down during the experiment and recorded using a minidisc recorder for further transcription and scoring. The total number of responses collected was 1408, with 352 times, 352 person names, and 704 times. For each category, half the responses are to the natural version, half to the synthesised version.

3.4.4. Scoring

Participants' pronunciations were scored by a phonetician (MW) based on whether their response was an acceptable pronunciation of the orthographical form of the target. This allows us to adjust for effects of the participants' dialect, such as rhoticity or differences in vowel quality. Deviant pronunciations that could not be accounted for by dialect were classified into three categories:

phoneme errors: Insertion, deletion or replacement of one consonant or vowel in a syllable. Example: Propanodryl → Prop**r**anodryl, Beclotor → Beclo**d**or. Phoneme errors occur in person names and medication names.

syllable errors: More than one phoneme error in the pronunciation of a syllable. Syllable errors only occur in medication names. Example: Propanodryl → Propano**lol**

word errors: One of the target words is replaced by a different word. Medication names were scored as wrong words if all of the word's syllables were affected by syllable errors. Word errors occurred in all three stimulus categories. Example: eleven → **seven**

Responses were scored as **correct** if they contained no errors.

Table 2: *Target Stimuli*

Item type	Items
Person	Ben, Bob, Dan, Don, Dick, Ned, Nick, Rick, Rob, Ron, Ken, Kim, Jim, Tim, Ted, Tom
Medication	Accumycin, Beclotor, Dexozine, Erytozole, Fosinarol, Kisinolol, Levapril, Mevacycline, Pravaclor, Propanodryl, Sulfacillin, Streptostatin, Tetradine, Trovalide
Times	one, four, five, seven, nine, ten, eleven, twelve ten past ten, ten past three, ten past twelve, ten past two, ten to eight, ten to eleven, ten to one, ten to ten twenty past ten, twenty past three, twenty past twelve, twenty past two, twenty to eight, twenty to eleven, twenty to one, twenty to ten quarter past ten, quarter past three, quarter past twelve, quarter past two, quarter to eight, quarter to eleven, quarter to one, quarter to ten

4. Results

Results are presented in three stages. First, we examine whether some stimuli were more difficult to process than others and present results of a detailed inspection of the synthetic speech signals that caused particular problems (Section 4.1). Next, we

examine the effect of ageing. Instead of testing chronological age, we focus on measures of cognitive ability (Section 4.2) and hearing loss(Section 4.3, both of which are linked to ageing.

4.1. The Effect of the Stimuli

We determined the effects of three independent variables characterising the nature of the stimuli, category (person, time, or medication), voice (synthetic or human) and position in the reminder (first or second), on participants' ability to remember the stimulus correctly. A three-way ANOVA shows main effects of the category (df=2,F=278.66,p$<$0.00001), voice (df=1,F=26.66,p$<$0.0001) and position (df=1,F=5.58,p$<$0.05). Tukey's HSD post-hoc tests reveal that synthetic stimuli are more difficult to remember than those spoken by the natural voice, items in second place are easier to remember than items in first place, and persons and times are easier to remember than medications (cf. Table 3). This validates our decision to test all three types of responses. The reasons for this result are clear: Times and person names are frequent, familiar, and phonologically simple, whereas medication names are unfamiliar and phonologically complex. We also find a clear interaction between stimulus category and voice (df=2, F=33.06, p$<$0.0000001). Our post-hoc tests reveal that in fact, participants remember times and person names well *no matter what the voice*—it is the complex, unfamiliar medication names that make the difference: Performance doubles for the natural voice compared to the synthetic voice. Therefore, when messages are restricted to stimuli using familiar words in familiar contexts, older users may be able to cope perfectly well with modern synthetic voices.

Although average scores for person names and times are similar, performance on the two categories is not correlated (ρ=-0.09,df=42,p$>$0.5). Neither is there a correlation between the number of correct person names and the number of correct medication names (ρ=0.19,p$>$0.2), nor between the number of correct times and the number of correct medication names (ρ=0.13,p$>$0.4). If participants' performance for the three response categories is uncorrelated, then performance on each category is potentially determined by different factors.

For six targets, the performance difference between natural and synthetic versions was 30% or worse. These were the medication names "Accumycin", "Beclotor", "Erytozole", "Mevacycline", "Pravaclor", and "Sulfacillin". In two of these, "Accumycin" and "Sulfacillin", there are clear bad joins. The second syllable of "Accumycin" is often misheard as "clu" or "cru". This could be due to a bad join in the /m/ of "mycin", where a nasal with relatively weak intensity meets a nasalised /a/. Likewise, "Sulfacillin" is affected by a bad join in the first vowel /ʊ/, and "-lin" is rendered as /lɪnɪn/. As a consequence, 33% of participants misheard the suffix, and 50% confused the initial /ʊ/ with an initial /ɪ/. In the remaining four, "Erytozole", "Mevacycline", "Pravaclor", and "Beclotor", the problem lies elsewhere. With "Mevacycline", "meva-" is often misheard as "neva-". This could be due to a tricky transition between the final /r/ of "your" and the initial /m/ of "meva". With "Pravaclor", the third syllable is affected most, with participants omitting the /l/, which is very short, or changing the nucleus to /a/, which may be due to the almost vocalic final /r/. For "Erytozole", the suffix "zole" is often confused with a similar sounding suffix. This could be due to the relatively rapid transition to the following preposition "at". "Beclotor" was affected worst. This is not due to bad joins, but to very short nuclei whose identity is difficult to identify. Moreover, the final /r/ is very short and segues quickly into the initial vowel of the following "at". As a result, none of the participants identifies the suffix correctly. Most misinterpret "-tor" as "-tin", and only seven correctly identify the /l/ in "-clo".

The picture sketched above for the six medications with the biggest performance difference between the natural and the synthetic version holds for the other medications as well: Bad joins are less of a problem than transitions that are too fast and durations, in particular of second consonants in consonant clusters, that are too short.

Table 3: *% correct by voice and stimulus category*

Category	**Voice**		Total
	Natural	Synthetic	
Medication	65.91%	35.23%	50.57%
Person	96.59%	90.91%	93.75%
Time	94.60%	96.02%	95.31%
Total	87.93%	79.55%	83.75%

4.2. The Effect of Memory

Working memory score is highly correlated with participants' performance on natural stimuli (ρ=0.42, 95% confidence interval [0.14,0.64], p$<$0.01), but not with performance on synthetic stimuli in general (ρ=0.23, 95% CI [-0.07,0.49], p$>$0.1). Looking at the effect of working memory span on the kinds of errors made, we find a significant correlation with words substituted (ρ=-0.36, 95% CI [-0.59,-0.07], p$<$0.05), but not with altered phonemes or syllables.

4.3. The Effect of Hearing

After examining potential confounders such as particularly difficult items and working memory, we turn to the central aspect of our study, the influence of hearing. We are looking for aspects of hearing that are highly correlated with participants' performance: the number of correct responses, the amount of phoneme errors, the number of syllable errors, and the number of word errors. The audiological measures included in our analysis (cf. Sec. 3.3) are:

Pure Tone Audiometry: TRADL, TRADR, F2L, F2R, UHFL, UHFR

Central Auditory Processing: MAXR, MAXL (Speech audiometry); GAP (gap detection in noise)

It would be very convenient if most of the results obtained were due to participants with abnormal hearing that would have been eliminated automatically by the traditional screening test, with the average threshold TRAD for 0.5, 1, 2, and 4 kHz at 20dB or lower for both. For this reason, we present results for two groups of participants:

Full: the complete group of 44 participants

Screened: the subgroup of 35 (79.55%) participants who would have passed the traditional screening test

Of the group SCREENED, 5 (14%) had a gap detection threshold in noise of 20 ms or higher. 8 (23%) had to hear the speech audiometry word lists at 60dB or louder to obtain a perfect score. This is well above the dynamic range of normal speech, which varies between 20 and 50 dB.

Tables 4–7 summarise the audiological measures which correlate with participants' performance on synthetic versus

natural speech. Measures for which correlations are significant at a level of $p<0.005$ are presented in **bold**, correlations with $p<0.01$ are in normal type, and measures for which correlations are significant at $p<0.05$ in *italics*.

All correlations are in the expected direction: the higher audiometric thresholds, the higher the gap detection threshold, and the higher the maximum intensity at which participants correctly repeated all words, the worse their performance. The first key result to note is that for both full group and screened group, aspects of hearing clearly influence performance. This is a powerful argument for including at least some simple hearing thresholds as covariates when analysing results of intelligibility tests. Even though our population was significantly older than the usual undergraduate testers, age does not imply healthy ears: We excluded two younger subjects from our initial pool of 15 younger participants because of low-frequency hearing loss.

The key *hearing threshold* is not one of the traditional screening values TRADR and TRADL, but F2L. This is the one threshold that correlates well with our error measures, no matter what the group. This is good news, because like TRADL, it is relatively quick and easy to measure. We also find strong correlations with the ultra-high frequency hearing thresholds UHFR and UHFL, which confirms our earlier findings [9]. The correlations between UHFR and UHFL and participant performance are stronger for the subgroup that would have passed screening than for the full group. This is interesting, since losses at ultra-high frequencies precede losses further down the basilar membrane.

Our measurements of *central auditory function*, MAXR, MAXL, and GAP are mainly correlated with participants' performance on natural speech - they play a far smaller role in predicting errors on synthetic speech. In particular, MAXR correlates well with the number of correct responses, and the number of word errors. This reflects the design of this particular test, which looks at the ability to correctly understand monosyllables. GAP is only relevant in accounting for syllable errors made when repeating synthetic stimuli (cf. Table 6): The less participants are able to detect small gaps in noise, the more likely they are elide, substitute, or insert two or more phonemes in a syllable of a complex multisyllabic stimulus.

Finally, the evidence shows very clearly that hearing problems affect natural and synthetic speech differently, even though the underlying speaker was the same. The key differences are:

- Speech audiometry correlates far better with people's ability to understand natural speech than with their ability to understand synthetic speech.
- Performance for synthetic stimuli on the other hand is predicted mostly by pure tone audiometry thresholds.
- No audiological measures correlate significantly with the number of phoneme errors made on synthetic speech, and no measures correlate significantly with the number of syllable errors made on natural speech.

Table 4: *Correlation of audiological measures with performance on reminder task*

	Full (n=44)	
	Natural	Synthetic
Audiometry	**F2L**	F2L, UHFR
	TradL	*UHFL, TradL*
Central	**MaxR**, MaxL	MaxL
	Screened (n=35)	
	Natural	Synthetic
Audiometry	(none)	**F2L, UHFR, UHFL**
		F2R, TradL, *TradR*
Central	(none)	(none)

Table 5: *Correlation of audiological measures with phoneme errors*

	Full (n=44)	
	Natural	Synthetic
Audiometry	**F2L, TradL**	(none)
	F2R, TradR, UHFL, UHFR	
Central	*MaxR*	(none)
	Screened (n=35)	
	Natural	Synthetic
Audiometry	*TradR, F2R*	(none)
Central	(none)	(none)

5. Discussion

Our results indicate that older people can remember and process synthetic stimuli just as well as those produced by natural speech if the text consists of familiar words and phrases. We can exploit this finding by ensuring that prompts are redundant and contain frequent and familiar words. Since quite a few problems with the synthetic stimuli occurred at transitions between the target words and the surrounding sentence matrix, a quick hack to avoid these problems would be to delimit the key content words by very short pauses. These general design guidelines can be implemented almost immediately and benefit all users regardless of age.

Considerable differences emerge only when the text to be synthesised contains phonologically complex, unfamiliar stimuli. This result needs to be investigated further in a more systematic study where phonological complexity and familiarity are both varied systematically.

Our results also demonstrate that factors which will affect the ability to understand natural speech do not necessarily affect the ability to understand synthetic speech. Hence, we cannot just extrapolate from the literature on human speech recognition, but need to reevaluate all findings carefully.

A more detailed analysis of the results shows that people's ability to understand synthetic speech is greatly influenced by pure-tone audiometric thresholds. Central auditory processing has a small, but decisive influence. For example, when remembering phonologically complex syllables, the ability to detect small gaps in the signal becomes important. This indicates that users' ability to understand synthetic speech may depend mainly on aspects of auditory function that affect the general processing of auditory stimuli, and less on users' ability to understand speech.

The natural response to this result might be to apply preemphasis to relevant frequency ranges. However, the benefits of any signal processing need to be weighed against the distortions it introduces. Furthermore, detailed post-hoc error analyses show that the main source of errors are not bad joins, but segments that are too short and transitions that move too quickly. Hence, it might be more effective to use units for important content words that are longer and contain clearer auditory cues. We hope to investigate this hypothesis in future work.

Table 6: *Correlation of audiological measures with syllable errors*

	Full (n=44)	
	Natural	Synthetic
Audiometry	(none)	F2L, F2R, *UHFL, UHFR*
Central	(none)	**MaxR**, *Gap*
	Screened (n=35)	
	Natural	Synthetic
Audiometry	(none)	*F2L, UHFR, UHFL*
Central	(none)	*Gap*, *MaxL*

Table 7: *Correlation of audiological measures with word errors*

	Full (n=44)	
	Natural	Synthetic
Audiometry	(none)	*F2L, TradL*
Central	*MaxR*	(none)
	Screened (n=35)	
	Natural	Synthetic
Audiometry	(none)	TradR, *TradL, UHFL UHFR, F2R, F2L*
Central	MaxR, *MaxL,GapNoise*	(none)

6. Acknowledgements

This research was funded by the EPSRC/BBSRC initiative SPARC and by the SFC grant MATCH (grant no. HR04016). We would like to thank our participants for their patience, Matthew Aylett and Christopher Pidcock for their invaluable help with generating the stimuli, and Ravi Chander Vipperla for his generous help with digitising minidiscs.

7. References

[1] B. Langner and A. W. Black, "Using Speech In Noise to Improve Understandability for Elderly Listeners," in *Proceedings of ASRU, San Juan, Puerto Rico*, 2005.

[2] R. W. Roring, F. G. Hines, and N. Charness, "Age differences in identifying words in synthetic speech," *Hum Factors*, vol. 49, pp. 25–31, 2007.

[3] S. Duffy and D. Pisoni, "Comprehension of synthetic speech produced by rule: A review and theoretical interpretation," *Language and Speech*, vol. 35, pp. 351–389, 1992.

[4] C. R. Paris, M. H. Thomas, R. D. Gilson, and J. P. Kincaid, "Linguistic cues and memory for synthetic and natural speech," *Human Factors*, vol. 42, pp. 421–431, 2000.

[5] P. Luce, T. Feustel, and D. Pisoni, "Capacity demands in short-term memory for synthetic and nautral speech," *Human Factors*, vol. 25, pp. 17–32, 1983.

[6] J. Al-Awar Smither, "The processing of synthetic speech by older and younger adults," in *Proceedings of the Human Factors Society 36th Annual Meeting. Innovations for Interactions, 12-16 Oct. 1992*. Atlanta, GA, USA: Human Factors Soc, 1992, pp. 190–192.

[7] A. Black and P. Taylor, "The festival speech synthesis system," Human Communication Research Centre, Tech. Rep. TR-83, 1997.

[8] B. Sutton, J. King, K. Hux, and D. Beukelman, "Younger and older adults' rate performance when listening to synthetic speech," *Augmentative and Alternative Communication*, vol. 11, no. 3, pp. 147–153, 1995.

[9] M. Wolters, P. Campbell, C. DePlacido, A. Liddell, and D. Owens, "The effect of hearing loss on the intelligibility of synthetic speech," in *Proc. Intl. Conf. Phon. Sci.*, Aug. 2007.

[10] T. A. Salthouse, "Where in an ordered sequence of variables do independent age-related effects occur?" *J.Gerontol.B Psychol.Sci.Soc.Sci.*, vol. 51, pp. 166–178, 1996.

[11] J. R. Crawford, G. Smith, E. A. Maylor, S. della Sala, and R. H. Logie, "The Prospective and Retrospective Memory Questionnaire (PRMQ): Normative data and latent structure in a large non-clinical sample," *Memory*, vol. 11, pp. 261–275, ,Psychological physiopathology 2003.

[12] N. Unsworth and R. Engle, "Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects," *Journal of Memory and Language*, vol. 54, pp. 68–80, 2006.

[13] P. Rabbitt, "Mild hearing loss can cause apparent memory failures which increase with age and reduce with IQ," *Acta Otolaryngol.Suppl*, vol. 476, pp. 167–175, 1990.

[14] British Society of Audiology, "Pure tone air and bone conduction threshold audiometry with and without masking and determination of uncomfortable loudness levels," 2004.

[15] R. W. Keith, *The Random Gap Detection Test.* St. Louis: AUDITEC, 2000.

[16] D. Owens, P. Campbell, A. Liddell, C. DePlacido, and M. Wolters, "Random gap detection threshold: A useful measure of auditory ageing?" in *Proc. Europ. Cong. Fed. Audiol. Heidelberg, Germany*, Jun.

[17] K. B. Snell, F. M. Mapes, E. D. Hickman, and D. R. Frisina, "Word recognition in competing babble and the effects of age, temporal processing, and absolute sensitivity," *Journal of the Acoustical Society of America*, vol. 112, pp. 720–727, 2002.

[18] A. Boothroyd, "Developments in speech audiometry," *British Journal of Audiometry*, vol. 2, pp. 3–10, 1968.

[19] M. Pollack, "Intelligent Technology for an Aging Population: The Use of AI to Assist Elders with Cognitive Impairment," *AI Magazine*, vol. 26, pp. 9–24, 2005.

[20] J. R. Dubno and H. Levitt, "Predicting Consonant Confusions from Acoustic Analysis," *Journal of the Acoustical Society of America*, vol. 69, pp. 249–261, 1981.

[21] S. Gordon-Salant, "Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing," *J.Acoust.Soc.Am.*, vol. 80, pp. 1599–1607, 1986.

[22] M. A. Aylett, C. J. Pidcock, and M. E. Fraser, "The cerevoice blizzard entry 2006: A prototype database unit selection engine," in *Proceedings of Blizzard Challenge Workshop, Pittsburgh, PA*, 2006.